Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling heterogeneous variance–Covariance components in two-level models. *Journal of Educational and Behavioral Statistics*, *39*(5), 307-332. https://doi.org/10.3102/1076998614546494

Modeling Heterogeneous Variance-Covariance Components in Two-Level Models

George Leckie
Senior Lecturer in Social Statistics
Centre for Multilevel Modelling
Graduate School of Education
University of Bristol
2 Priory Road, Bristol
BS8 1TX
United Kingdom
E-mail: g.leckie@bristol.ac.uk
Telephone: +44 0117 33 10614

Robert French
Research Associate
Centre for Multilevel Modelling
Graduate School of Education
University of Bristol
2 Priory Road, Bristol
BS8 1TX
United Kingdom

Chris Charlton
Senior Software Engineer
Centre for Multilevel Modelling
Graduate School of Education
University of Bristol
2 Priory Road, Bristol
BS8 1TX
United Kingdom

William Browne
Professor of Biostatistics
Centre for Multilevel Modelling
School of Veterinary Science
University of Bristol
Langford, Bristol
BS40 5DU
United Kingdom

Authors

GEORGE LECKIE is a Senior Lecturer in Social Statistics at the Centre for Multilevel Modelling, Graduate School of Education, University of Bristol, 2 Priory Road, Bristol, BS8 1TX, United Kingdom; e-mail: g.leckie@bristol.ac.uk. His research interests are in the application of multilevel and other latent variable models for analyzing complex structured social science data, especially in educational research.

ROBERT FRENCH is a Research Associate at the Centre for Multilevel Modelling, Graduate School of Education, University of Bristol, 2 Priory Road, Bristol, BS8 1TX, United Kingdom; e-mail: r.french@bristol.ac.uk. His research interests are quantitative educational research, especially school effectiveness.

CHRIS CHARLTON is a Senior Software Engineer at the Centre for Multilevel Modelling, Graduate School of Education, University of Bristol, 2 Priory Road, Bristol, BS8 1TX, United Kingdom; e-mail: c.charlton@bristol.ac.uk. His research interests are in statistical software development.

WILLIAM BROWNE is a Professor of Biostatistics in the school of Veterinary Science, University of Bristol, Langford, Bristol, BS40 5DU, UK and director of the Centre for Multilevel Modelling; e-mail: william.browne@bristol.ac.uk. His research interests are in the use of statistical modeling techniques, in particular MCMC methods, for analyzing complex datasets in many fields including veterinary epidemiology, ecology and education.

Acknowledgments
This research was funded under the LEMMA3 project, a node of the UK Economic and Social Research Council's National Centre for Research Methods (grant number RES-576-25-0035). We are very grateful to the helpful comments made by the four reviewers.

Modeling Heterogeneous Variance-Covariance Components in Two-Level Models

Abstract

Applications of multilevel models to continuous-outcomes nearly always assume constant residual variance and constant random-effects variances and covariances. However, modeling heterogeneity of variance can prove a useful indicator of model misspecification and in some educational and behavioral studies it may even be of direct substantive interest. The purpose of this paper is to review, describe and illustrate a set of recent extensions to two-level models that allow the residual and random-effects variance-covariance components to be specified as functions of predictors. These predictors can then be entered with random-coefficients to allow the level-1 heteroskedastic relationships to vary across level-2 units. We demonstrate by simulation that ignoring level-2 variability in residual variances leads the level-1 variance-function regression coefficients to be estimated with spurious precision. We discuss software options for fitting these extensions and we illustrate them by reanalyzing the classic High School and Beyond data and two-level school effects models presented by Raudenbush and Bryk (2002).

Keywords: *heterogeneous within-group variances*; *heteroskedasticity; log-linear variance models; multilevel models; variance-functions*

1. Introduction

In school effects research, two-level students-within-schools models are regularly fitted to continuous student achievement outcomes assuming constant variance-covariance components. Interest lies in measuring the effects of school organization, policy and practice on student mean achievement (e.g., Aitkin and Longford, 1986; Goldstein et al., 1993; Raudenbush and Bryk, 1986). However, different school processes may quite plausibly also produce important differences in variability in student achievement (Raudenbush and Bryk, 1987). For example, a new educational program shown to increase mean achievement may actually be undesirable if it simultaneously increases achievement dispersion to the extent that the number of students failing some minimum level of achievement increases. As well as varying between schools, achievement dispersion may also vary within schools, for example, as a function of student gender with, say, boys' tending to score more variably than girls (Browne et al., 2002; Goldstein and Thomas, 1996). Such a finding would also be substantively important as it would show boys' future performances are less predictable than those for girls (Snijders and Bosker, 2012).

The purpose of this paper is to review, describe and illustrate a set of recent extensions that have been proposed in the literature for modeling and testing hypotheses about the sources of variance-covariance heterogeneity in multilevel models (e.g., Hedeker et al., 2008; Lee and Nelder, 2006). These methods extend the general two-level model (Goldstein, 2011; Snijders and Bosker, 2012; Raudenbush and Bryk, 2002) by modeling the residual variance as a log-linear function of level-1 and level-2 predictors and by allowing the level-1 predictors to enter this variance-function with random coefficients. The level-2

variances and correlations (standardized covariances) can then be further modeled as log-linear and inverse-tanh-linear functions of the level-2 predictors.

Raudenbush and Bryk (2002) view residual variance heterogeneity as an omnibus signal of model misspecification, indicating a need for a richer level-1 mean-function. An omitted level-1 predictor with unequal variance across level-2 units will tend to give rise to heterogeneity of level-1 variance across level-2 units. An included level-1 predictor treated erroneously as fixed when it should be treated as random or nonrandomly varying (i.e., omitted random-coefficients or omitted cross-level interactions) will tend to lead the level-1 variance to be a function of that predictor.

Raudenbush and Bryk (1987) discuss variance heterogeneity in the context of measuring school program effects on achievement dispersion (see also Kim and Seltzer, 2011). When schools are randomly assigned to treatments, heterogeneity of residual variance across treatments provides evidence of omitted interactions between treatments and unspecified student characteristics (i.e., student differences in response to treatments). They recommend that these characteristics are identified and included in the model. Ideally, heterogeneity of variance will then disappear. Typically it is not possible to identify all relevant characteristics (e.g., due to insufficient data) in which case a significant treatment dispersion difference serves as a warning that unmodeled treatment interaction effects remain. In non-randomized settings, variance heterogeneity across treatments may additionally reflect variance heterogeneity in unmodeled student background characteristics across these schools. It is therefore important to enter into the model not only those student-level characteristics predictive of the outcome whose means differ across treatments (in order to measure mean achievement differences appropriately), but

to additionally include those characteristics which are predictive of the outcome whose variances are unequal across treatments. A further issue in non-randomized settings is that variance heterogeneity across treatments may also reflect other school-characteristics associated with the selection of schools into treatment, including important context variables (variables describing the composition of the student body). Finally, heterogeneity of variance may also indicate floor or ceiling effects in the achievement scale.

To test hypotheses about the sources of level-1 heterogeneity, the mean and variance-functions should be modeled jointly (Aitkin, 1999; Lee and Nelder, 1996, 2001). The significance of the variance-function regression coefficients can then be assessed via Wald tests in the usual way. It is well known that ignoring clustering in linear regression leads the regression coefficients to be estimated with spurious precision, especially regression coefficients relating to cluster-level predictors. A secondary purpose of this paper is to show via simulation that a parallel argument applies when modeling the residual variance as a function of predictors in two-level models. We are not aware of this point being made before in the literature.

In the next section, we review extensions proposed by methodologists for modeling heterogeneous variance-covariance components in two-level models, particularly approaches to include random effects in level-1 variance-functions. Section 3 combines these extensions into a general model and summarizes software estimation options. Section 4 presents the simulation study. Section 5 illustrates the modeling extensions by reanalyzing the classic High School and Beyond (HSB) data (Raudenbush and Bryk, 2002) and by addressing the following example research question: do Catholic schools produce higher mean achievement and narrower achievement dispersion than Public schools, even

after adjusting for school differences in student background? We conclude with a discussion in Section 6.

2. Review

Several multilevel textbooks discuss modeling variance components as functions of predictors. Raudenbush and Bryk (2002, page 131) discuss modeling the level-1 variance as a log-linear function of predictors, an approach implemented in their HLM software (Raudenbush et al., 2012), and long applied in modeling heteroskedasticity in single-level linear models (Aitkin, 1987; Davidian & Carroll, 1987; Harvey, 1976). Goldstein (2011, Sections 3.1, 3.2) and Snijders and Bosker (2012, Chapter 8) discuss modeling the level-1 variance and level-2 variances as linear functions of the predictors, an approach implemented in the MLwiN software (Rasbash et al., 2009). However, Goldstein (2011, Section 9.4) ultimately recommends specifying log- rather than identity-link functions for the variance components to ensure positive values.

In school effects research, several papers have modeled the level-1 variance not only as a function of predictors, but as varying randomly across level-2 units. Raudenbush and Bryk (1987) propose a two-step approach where they first fit a standard two-level random-intercept model to student achievement assuming constant within-school variances. They then fit a single-level linear model to regress the log of the within-school variances of the step one residuals on school characteristics. While straightforward to implement, this approach precludes student-level predictors from the step two model and does not propagate the uncertainty in estimating the step one parameters into the step two model. Kasim and Raudenbush (1998) extend the standard two-level random-intercept

model by jointly modeling the within-school variances as inverse chi-squared distributed. They show, through simulation, that the mean-function parameters are largely robust to naively assuming a constant level-1 variance in the presence of random heterogeneity across level-2 units. Kim and Choi (2008) extend the standard two-level random-coefficient model for student achievement by jointly modeling the square root of the within-school variance as a linear function of both a school-level predictor and a normally distributed random-intercept effect. They also allow an association between the mean and within-school variance by entering the mean-function random-intercept as a latent predictor in the variance-function.

Researchers in other fields have also started to model the level-1 variance as randomly varying across level-2 units. In applied statistics, the double hierarchical generalized linear model (DHGLM) framework proposed by Lee and Nelder (2006) allows the level-1 variance to be modeled as a log-linear function of predictors and a level-2 random-intercept effect (see also, Lee et al., 2006). In psychology, Hedeker et al. (2008) model both the level-1 and level-2 variances in a two-level observations-within-subjects random-intercept model as a log-linear function of predictors. A random-intercept effect is included in the level-1 variance-function and is assumed correlated with the usual mean-function random-intercept effect. Hedeker and Mermelstein (2012) extend the mean-function in this model to include random coefficients, but restrict the level-2 covariance matrix to be constant. Jahng (2008) and Rast et al. (2012) also present two-level random-coefficient models, but they allow the effects of the level-1 predictors in their level-1 variance function to also vary randomly across level-2 units. They specify constant level-2 covariance matrices.

3. Methods

The standard two-level model

First consider the standard two-level model for continuous-outcome Y_{ij} and level-1 and level-2 predictor variables X_{ij} and W_j, where i ($i = 1,2, \dots, n_j$) indexes the level-1 units and j ($j = 1,2, \dots, J$) indexes the level-2 units. The model, expressed using the notation and hierarchical form popularized by Raudenbush and Bryk (2002), is written as

$$\text{Level 1:} \qquad Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}, \tag{1.1}$$

$$\text{Level 2:} \qquad \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}, \tag{1.2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}, \tag{1.3}$$

$$\text{Combined:} \quad Y_{ij} = \underbrace{\gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij}}_{\text{fixed part}} + \underbrace{u_{0j} + u_{1j}X_{ij} + r_{ij}}_{\text{random part}}. \tag{1.4}$$

The level-1 model (1.1) regresses Y_{ij} on X_{ij}, specifying a separate intercept β_{0j} and slope β_{1j} for each level-2 unit. The level-1 residual r_{ij} is assumed normally distributed with zero mean and constant variance σ^2 across the level-1 units, that is $r_{ij} \sim \mathrm{N}(0, \sigma^2)$. The level-2 models, (1.2) and (1.3), regress the level-1 coefficients, β_{0j} and β_{1j}, on the level-2 predictor W_j. The level-2 random effects, u_{0j} and u_{1j}, are assumed bivariate normally distributed with constant covariance matrix across the level-2 units

$$\mathbf{T}_u = \begin{bmatrix} \tau_{u00} & \\ \tau_{u01} & \tau_{u11} \end{bmatrix}.$$

Substituting 1.2 and 1.3 into 1.1 gives the combined form (mixed-effects formulation or reduced form) of the model (1.4).

Modeling heterogeneous level-1 variances

The constant level-1 variance assumption can be relaxed by modeling it as a log-linear function of the level-1 and level-2 predictors and associated random effects (e.g., Hedeker et al., 2008; Rast et al., 2012)

Level 1: $$\log(\sigma_{ij}^2) = \alpha_{0j} + \alpha_{1j}X_{ij}, \quad (2.1)$$

Level 2: $$\alpha_{0j} = \delta_{00} + \delta_{01}W_j + v_{0j}, \quad (2.2)$$

$$\alpha_{1j} = \delta_{10} + \delta_{11}W_j + v_{1j}, \quad (2.3)$$

Combined: $$\log(\sigma_{ij}^2) = \underbrace{\delta_{00} + \delta_{01}W_j + \delta_{10}X_{ij} + \delta_{11}W_jX_{ij}}_{\text{fixed part}} + \underbrace{v_{0j} + v_{1j}X_{ij}}_{\text{random part}}. \quad (2.4)$$

The level-1 model (2.1) regresses the log of σ_{ij}^2 on X_{ij}, specifying a separate intercept α_{0j} and slope α_{1j} for each level-2 unit. The level-2 models, (2.2) and (2.3), then regress α_{0j} on W_j and α_{1j} on W_j, respectively. Note that while (2.1), (2.2) and (2.3) are specified in terms of the same predictors as in (1.1), (1.2) and (1.3), this is not a requirement of the model. The level-2 random effects, v_{0j} and v_{1j}, are assumed bivariate normally distributed with constant covariance matrix across the level-2 units, that is

$$\mathbf{T}_v = \begin{bmatrix} \tau_{v00} & \\ \tau_{v01} & \tau_{v11} \end{bmatrix}.$$

When $\alpha_{0j} = \delta_{00}$ and $\alpha_{1j} = 0$, the level-1 variances are once again constant across level-1 units. Substituting 2.2 and 2.3 into 2.1 gives the combined form of the level-1 variance-function (2.4).

Association between the mean and level-1 variance-functions

Independence between the mean-function random effects, u_{0j} and u_{1j}, and the level-1 variance-function random effects, v_{0j} and v_{1j}, can be relaxed by modeling them as multivariate normally distributed (e.g., Rast et al., 2012) with a 4×4 level-2 covariance matrix

$$\mathbf{T}_{uv} = \begin{bmatrix} \tau_{u00} & & & \\ \tau_{u01} & \tau_{u11} & & \\ \tau_{u0v0} & \tau_{u1v0} & \tau_{v00} & \\ \tau_{u0v1} & \tau_{u1v1} & \tau_{v01} & \tau_{v11} \end{bmatrix}.$$

An alternative approach to inducing an association between the school means and level-1 variances would be to continue to model the two sets of random effects as independent, but to enter the mean-function random effects into the level-1 variance-function as latent covariates with regression coefficients to be estimated (e.g., Kim and Choi, 2008). By entering the mean-function random effects non-linearly, this approach can be extended to account for the concave relationship between the variance and the mean that would be expected when an outcome exhibits floor and ceiling effects.

Simulation studies have shown that inference for the general two-level model is relatively robust to violation of the normality assumption (e.g., McCulloch and Neuhaus,

2011). However, less is known about whether inferences in models with random-effects in the level-1 variance function are similarly robust. Where outlying level-2 units are of concern, it would be prudent to study the sensitivity of results to the level-2 (multivariate) normality assumptions, for example, by reanalyzing the data with heavy-tailed (multivariate) t-distributions (Seltzer et al., 1996). More generally, to safeguard against distributional misspecification, the (multivariate) random-effects distribution could in principle be left unspecified by using discrete random effects (i.e., latent classes) with the number of classes being specified by model fit criteria.

Modeling heterogeneous level-2 variance-covariance components

A further extension to the model is to treat the individual level-2 variances and correlations (standardized covariances) themselves as functions of the level-2 predictors. Consider, for simplicity, the special case of only a random-intercept in the mean-function and in the level-1 variance-function. The level-2 covariance matrix is then 2×2 with variances τ_{u00} and τ_{v00}, covariance τ_{u0v0} and correlation $\rho_{u0v0} = \tau_{u0v0}/\sqrt{\tau_{u00}\tau_{v00}}$. The variance- and correlation-functions for these three parameters can be written as

Level 2:
$$\log\left(\tau_{u00j}\right) = \kappa_{u000} + \kappa_{u001}W_j, \tag{3.1}$$
$$\log\left(\tau_{v00j}\right) = \kappa_{v000} + \kappa_{v001}W_j, \tag{3.2}$$
$$\tanh^{-1}\left(\rho_{u0v0j}\right) = \kappa_{u0v00} + \kappa_{u0v01}W_j, \tag{3.3}$$

where log link functions are specified to ensure positive variances and an inverse tanh link function to ensure the correlation ρ_{u0v0j} lies between -1 and $+1$ (c.f., Hedeker et al., 2008,

who specify a log-linear function for τ_{u00j}, but restrict $\tau_{v00j} = \tau_{v00}$ and $\tau_{u0v0j} = \tau_{u0v0}$). Note, that when variance-covariance matrices are of order three or larger, specifying appropriate link functions for the individual variances and correlations is a necessary but not sufficient condition to ensure that the matrix is positive definite. When $\kappa_{u001} = \kappa_{v001} = \kappa_{u0v01} = 0$, the level-2 variance-covariance matrices are once again constant across level-2 units.

Software and estimation

Restricted versions of the model with no random effects entering the level-1 variance-function can be fitted in SAS (SAS Institute Inc., 2013) using the PROC MIXED procedure (Littell, 2006, Chapter 9), or with dedicated multilevel modeling packages such as HLM (Raudenbush et al., 2012) or MLwiN (Rasbash et al., 2009). The two-level random-intercept version of the model with log-linear variance-functions at both level-1 and at level-2 and with a random-intercept effect entering the level-1 variance-function can be fitted in SAS PROC NLMIXED or in the stand-alone MIXREGLS package (Hedeker and Nordgren, 2013). The latter can be called from R (R Core Team, 2014) (via the mixregls_function.R function file; Hedeker and Nordgren, 2013) and Stata (StataCorp, 2013) (via the runmixregls command; Leckie, 2014). It can also be fitted using the DHGLM framework as implemented in ASReml (Gilmour et al., 2009), GenStat (Payne et al., 2009), and R (via the dhglm package, Noh and Lee, 2013), but in this framework the mean-function and level-1 variance-function random-intercepts are assumed independent.

The full model, with the addition of appropriate prior distributions, can be fitted using Markov chain Monte Carlo (MCMC) methods in several packages, the most prominent of

which is WinBUGS (Lunn et al., 2000). Rast et al. (2012) present WinBUGS syntax for a version of the model presented here which assumes constant level-2 variance-covariance components. We implement the full model using the e-Stat estimation engine within the Stat-JR (pronounced "stature") statistics package (Charlton et al., 2013) developed at the Centre for Multilevel Modelling (CMM).

We specify diffuse (vague, flat or minimally informative) prior distributions for all parameters. We specify Gaussian priors with zero means and very large variances (effectively, improper Uniform(-∞,∞) priors) for the regression coefficients. We specify an inverse-Wishart prior Wishart$^{-1}(\mathbf{R}_{uv}, n)$ for the level-2 covariance matrix $\mathbf{T}_{uv}$ where $\mathbf{R}_{uv}$ is the "scale matrix" and n the sample size on which the prior is based. We set n to be as small as possible (i.e., equal to the order of $\mathbf{T}_{uv}$) so that the prior is minimally informative. We set $\mathbf{R}_{uv} = n \times \widehat{\mathbf{T}}_{uv}$ where $\widehat{\mathbf{T}}_{uv}$ is some plausible initial estimate for $\mathbf{T}_{uv}$; we set the diagonal elements to values with the right order of magnitude and the off-diagonal elements to zero. One way to obtain these initial estimates is to fit the model separately to each school and then calculate the variability across schools in those coefficients modeled as random in the joint model. It is prudent to study the sensitivity of one's results to reasonable alternative specifications of the priors, especially for the level-2 covariance matrix when the number of level-2 units is small. Moderate changes to $\widehat{\mathbf{T}}_{uv}$ do not appreciably change the results in our illustration, but larger differences in posterior means and intervals would be expected in applications with fewer level-2 units. Specifying instead a uniform prior for $\mathbf{T}_{uv}$ gives slightly larger posterior means and wider posterior intervals. Should one instead wish to relax the constant level-2 covariance matrix assumption, proposed values for the parameters of the resulting variance and correlation-functions are only accepted when the

ensuing covariance matrix remains positive definite (as demonstrated in Browne, 2006). See Browne and Draper (2000) and Seltzer et al. (1996) for further discussion regarding sensitivity of model results to choice of priors.

We fit all models using five chains with dispersed starting values, each with a burn-in period of 20,000 iterations and a monitoring period of 20,000 iterations. Informal visual assessments of the parameter chains and standard MCMC convergence diagnostics suggest that these periods are sufficiently long. We visually inspect overlaid trace and density plots of the multiple chains to confirm they escape the influence of their starting values and converge to common stationary unimodal posterior distributions. We examine autocorrelation-function (ACF) plots and effective sample size statistics for different parameters as well as consulting standard MCMC convergence diagnostics to check that we run the MCMC sampler for sufficiently long. An examination of cross-correlations and bivariate plots gives no suggestion of overfitting problems (e.g., "ridges" where two parameters are nearly confounded). See Cowles and Carlin (1996) for further discussion regarding MCMC convergence diagnostics.

When presenting results, we report the means, SDs and 2.5th and 97.5th quantiles (95% credible intervals) of the 100,000 pooled monitoring iterations. These quantities are analogous to the parameter estimates, standard errors and lower and upper bounds of 95% confidence intervals obtained in a frequentist analysis. We use the deviance information criterion (DIC) to compare the fit of models (Spiegelhalter et al., 2002): models with smaller DIC values are preferred to those with larger values, with differences of five or more considered substantial (Lunn et al., 2012).

4. Simulation

We use the following data generating model to illustrate the consequences of ignoring level-2 variability in residual variances. The mean-function is written as

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij},$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j},$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j,$$

where u_{0j}~N$(0, \tau_{u00})$ and r_{ij}~N$\left(0, \sigma_{ij}^2\right)$. The level-1 variance-function is written as

$$\log\left(\sigma_{ij}^2\right) = \alpha_{0j} + \alpha_{1j}X_{ij},$$

$$\alpha_{0j} = \delta_{00} + \delta_{01}W_j + v_{0j},$$

$$\alpha_{1j} = \delta_{10} + \delta_{11}W_j,$$

where v_{0j}~N$(0, \tau_{v00})$. The predictors X_{ij} and W_j are standard normal variates (X_{ij} is simulated to have an intraclass correlation coefficient, ICC, of 0.2). We analyze 1,000 replications of 50 schools, 25 students per school.

We fit two models to the 1,000 replications. Model A incorrectly assumes that the level-1 variance is a deterministic function of the predictors ($\tau_{v00} = 0$). Model B matches the data generating process and estimates the intercept variance of the level-1 variance-function. Table 1 presents the averages and SDs of the parameter posterior means for these two models across the 1,000 replications. The table also presents the averages of the

parameter posterior SDs and the coverage percentages of the 95% posterior credible intervals.

The Model A and Model B averaged posterior means match their true values with the exception that the level-1 variance-function intercept δ_{00} is estimated to be -0.519 in Model A, compared to a true value, and that reported by Model B, of -0.611. This discrepancy is expected and relates to Model A's level-1 variance-function having a "population-averaged" interpretation, whereas Model B's level-1 variance-function has a "school-specific" interpretation. Specifically, $\exp(\delta_{00})$ in Model A measures the population-*averaged* within-school variance (when $X_{ij} = 0$ and $W_j = 0$), whereas $\exp(\delta_{00} + v_{0j})$ in Model B measures the within-school variance specific to school j. It can be shown that the population-average of the Model B variances is given by $\exp(\delta_{00} + \frac{1}{2}\tau_{v00})$, whereas simply calculating $\exp(\delta_{00})$ gives a lower value corresponding to the population-median within-school variance (see Appendix). The former is estimated to be $0.602 = \exp(-0.611 + \frac{1}{2} \times 0.207)$, effectively equal in magnitude to the Model A estimate, $0.595 = \exp(-0.519)$. Thus, in this simulation, ignoring unexplained variation in the level-1 variance across level-2 units does not lead to biased posterior means.

In contrast, the Model A averaged posterior SDs for the level-1 variance function parameters are substantially smaller than the SDs of the corresponding posterior means indicating that the parameter posterior SDs are biased downwards. The averaged posterior SD for δ_{01} (0.043) (the coefficient of the level-2 predictor) is approximately 48% smaller than the SD of its posterior mean (0.082), while the averaged posterior SDs for δ_{10} and δ_{11} (the coefficients of the level-1 predictor and cross-level interaction) are approximately 25% and 23% smaller than the SDs of their posterior means. The 95% credible interval for

δ_{01} includes the true value on only 71% of replications, while the 95% credible intervals for δ_{10} and δ_{11} include their true values only 86% of the time. Conversely, the Model B averaged posterior SDs for these parameters lie much closer to the SDs of their posterior means and their coverage percentages are very close to their nominal 95% value (94%, 94% and 95%, respectively).

Repeating the simulation study in a large data setting with 250 schools and 100 students per school increases the precision of all parameter estimates (results not presented). The Model A 95% credible interval for δ_{01} now includes the true value on only 46% of replications while the 95% credible intervals for δ_{10} and δ_{11} include their true values only 70% and 71% of the time, respectively.

In summary, ignoring unexplained variation in the level-1 variance across level-2 units leads the level-1 variance-function regression coefficients to be estimated with spurious precision, especially the regression coefficients of level-2 predictors.

5. Illustration

We illustrate the modeling extensions using the classic High School and Beyond (HSB) two-level students-within-schools data which provides the principal teaching example in the Hierarchical Linear Models text (Raudenbush and Bryk, 2002). We consider the following example research question: do Catholic schools produce higher mean achievement and narrower achievement dispersion than Public schools, even after adjusting for school differences in student background? Ultimately, if sufficient adjustments can be made for selection into schools then the adjusted school means and

dispersions would correspond to so-called "Type B" school effects (Raudenbush and Willms, 1995; Willms and Raudenbush, 1989) which aim to isolate the effects of schools' practices (e.g., administrative leadership, curricular context, utilization of resources, and classroom instruction) from schools' contexts (social and economic characteristics of the community in which the school is located and the demographic composition of the student body).

The data consist of 7,185 students (level-1 units) nested within 160 schools (level-2 units) (mean = 45 students per school, range = 14 to 67). The response is continuous student math achievement (mean = 12.748, SD = 6.878) and is denoted Y_{ij} for student i ($i = 1, \ldots, n_j$) in school j ($j = 1, \ldots, 160$). A histogram (not presented) suggests a slight ceiling effect (42 students are recorded achieving the highest attainable score despite the response taking 6032 unique values). One might consider transforming the response however this has to be balanced with interpretation of the model parameters and so we do not consider this here. The three predictor variables are: student socioeconomic status $(\text{SES})_{ij}$ (mean = 0.000, SD = 0.779); school average socioeconomic status $(\text{MEAN SES})_j$ (mean = 0.000, SD = 0.414); and school sector $(\text{SECTOR})_j$, a dichotomous variable coded zero for the 90 Public schools and one for the 70 Catholic schools.

Model 1

Our first model makes no attempt to adjust for differential selection into schools. The model simply estimates the raw sector differences in school means and dispersions by entering school sector into the mean-function and separately into the level-1 variance function. The mean function is

$$Y_{ij} = \beta_{0j} + r_{ij},$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{SECTOR}_j + u_{0j},$$

where $u_{0j} \sim \text{N}(0, \tau_{u00})$ and $r_{ij} \sim \text{N}(0, \sigma_j^2)$. The level-1 variance function is

$$\log(\sigma_j^2) = \alpha_{0j},$$

$$\alpha_{0j} = \delta_{00} + \delta_{01}\text{SECTOR}_j + v_{0j},$$

where $v_{0j} \sim \text{N}(0, \tau_{v00})$. We allow u_{0j} and v_{0j} to be associated with covariance τ_{u0v0}. The corresponding correlation is derived in the usual way, $\rho_{u0v0} = \sigma_{u0v0}/\sqrt{\tau_{u00}\tau_{v00}}$.

The results (Table 2) show a significant sector difference γ_{01} in school means. Catholic schools, on average, score 2.829 points (equivalent to $2.829/6.878 = 0.411$ of an SD) higher than Public schools. The sector difference in the log of the within-school variances δ_{01} is also significant and is estimated as -0.266, and so Catholic schools, on average, show less dispersed achievement than Public schools. Specifically, the Catholic sector population-averaged within-school variance is estimated as $\exp(\delta_{00} + \delta_{01} + \frac{1}{2}\tau_{v00}) = 33.990$, which is substantially lower than the corresponding Public sector variance, estimated as $\exp(\delta_{00} + \frac{1}{2}\tau_{v00}) = 44.331$.[1] The estimated negative correlation between the two sets of random-intercept effects is not significant and so could be removed from the model.

To test whether we can remove the dispersion random effects entirely, we compare Model 1 to a restricted model which omits the dispersion random effect (not presented).

The DIC statistic is $46830 - 46817 = 13$ points lower in Model 1, confirming that achievement dispersion does vary significantly across schools, even after adjusting for the overall sector difference in dispersion. We also see the pattern of spurious precision suggested by the simulation study: the posterior SD of the estimated sector dispersion difference, δ_{01}, is 0.034 in the model which ignores the heterogeneity of residual variance (not presented) compared to a value of 0.040 in the model which correctly accounts for it.

The ICC, given by $\rho_j = \tau_{u00}/\left(\tau_{u00} + \sigma_j^2\right)$, varies randomly across schools. An examination of these ICCs (not presented), reveals the schoolmate achievement correlations range from 0.120 to 0.207 and so students are considerably more alike in some schools than in others.

Quantile-quantile (Q-Q) plots (normal scores plots) of the random effects (Figure 1) suggest the normality assumptions are plausible, although there is a suggestion that the mean and level-1 variance-function random-intercept effects are both somewhat negatively skewed.

The model specifies a constant level-2 covariance matrix $\mathbf{T}_{uv}$ and therefore assumes the two sectors are equally diverse in terms of their schools' mean performances and their schools' dispersion performances. We can relax this assumption by specifying the variances and correlation (standardized covariance) of this matrix as functions of school sector (as in Equation 3 where W_j denotes school sector).

Fitting these functions lowers the DIC by $46817 - 46802 = 15$ points indicating an overall improvement in model fit and that the sectors are therefore differentially diverse.[2] The estimates of the existing model parameters are similar to before and so we do not

report them here. To help interpret the estimated new parameters, we derive the matrix for each sector.

Public sector

$$\mathbf{T}_{uv} = \begin{bmatrix} \tau_{u00} & \rho_{u0v0} \\ \tau_{u0v0} & \tau_{v00} \end{bmatrix} = \begin{bmatrix} 6.635 & 0.490 \\ 0.194 & 0.024 \end{bmatrix}$$

Catholic sector

$$\mathbf{T}_{uv} = \begin{bmatrix} \tau_{u00} & \rho_{u0v0} \\ \tau_{u0v0} & \tau_{v00} \end{bmatrix} = \begin{bmatrix} 6.783 & -0.650 \\ -0.311 & 0.035 \end{bmatrix}$$

The between-school variance τ_{u00} is estimated to be $6.783 - 6.635 = 0.148$ units higher for the Catholic sector than for the Public sector. This difference is however substantively quite small and not significant. This is in contrast to the population-averaged within-school variance $\exp\left(\delta_{00} + \delta_{01}\text{SECTOR}_j + \frac{1}{2}\tau_{v00}\right)$ where the Catholic sector estimate is $44.452 - 34.217 = 10.235$ units smaller than for the Public sector.

The variance of the adjusted log Catholic within-school variances is 0.035 compared to 0.024 among Public schools and so the Catholic sector is also more diverse in terms of the degree of achievement dispersion in its schools. The correlation between adjusted school means and their log achievement dispersion also differs across the two sectors. In the Public sector, schools which show higher mean achievement tend to show more dispersed achievement ($\rho_{u0v0} = 0.490$). In the Catholic sector, the opposite is the case ($\rho_{u0v0} = -0.650$). An inspection of the three sector differences, $\kappa_{u001}, \kappa_{v001}, \kappa_{u0v01}$, suggests that only the sector correlation difference, κ_{u0v01}, is individually significant. For simplicity and because sector differences in the diversity of mean and dispersion effects are not the focus of the posed research question, we return to specifying a constant matrix in subsequent models.

Model 2

A fundamental concern with Model 1 is that the Catholic sector might, on average, recruit students with higher prior achievement and other background characteristics associated with success and it is these intake differences which, at least in part, explain their superior performance. Ignoring these differences will lead us to overestimate any true difference in sector means (i.e., omitted variable bias). If the Catholic sector also recruits a more homogenous intake with respect to important determinants of achievement, then we will also overestimate any true sector dispersion difference. We have omitted SES from the model, but SES is positively associated with achievement ($r = 0.361$). SES is also higher, on average, in Catholic schools (mean $= 0.150$) than Public schools (mean $= -0.146$) (a difference of 0.380 of an SD), and is somewhat less variable in Catholic schools (SD $= 0.741$) than in Public schools (SD $= 0.788$). Thus, the higher mean achievement and lower dispersion seen in Catholic schools may be a direct consequence of schools' practices and preferences in the recruitment of students with respect to SES.

We shall enter SES into the mean-function to attempt to adjust for these potential selection effects. In doing so, it is important to specify the within-school achievement-SES relationship correctly, otherwise part of the "true" association will be absorbed into the residual. This may bias the adjusted sector dispersion difference and artificially make the residual variance a function of SES. Graphical exploration of the data suggests: (1) the social gradient in achievement varies from school to school (a random-slope on SES is required); (2) the social gradient in achievement is in general stronger in Public schools than in Catholic schools (a cross-level interaction between sector and SES is required); (3) the social gradient in achievement is stronger between schools compared to within schools

(we need to specify separate within- and between-school effects). Model 2 attempts to capture these different features by specifying the following mean-function

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{SES} - \text{MEAN SES})_{ij} + r_{ij},$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{SECTOR}_j + \gamma_{02}(\text{MEAN SES})_j + u_{0j},$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\text{SECTOR}_j + \gamma_{12}(\text{MEAN SES})_j + u_{1j},$$

where u_{0j} and u_{1j} are assumed bivariate normal with zero mean and constant covariance matrix and $r_{ij} \sim \text{N}(0, \sigma^2)$.

The results (Table 2) show the six fixed effects are significantly different from zero. As expected, the adjusted sector mean and dispersion differences are now substantially smaller than their raw counterparts. The estimated sector mean difference γ_{01} when holding SES constant, is estimated as 1.258, compared to a raw difference of 2.829. The other fixed effects confirm that schools with higher mean SES have higher mean achievement ($\gamma_{02} = 5.281$) and that the positive within-school achievement-SES relationship is significantly stronger in Public schools than in Catholic schools ($\gamma_{11} = -1.664$) and is stronger in high-SES schools than in low SES schools ($\gamma_{12} = 1.045$).

The sector difference in the log of the within-school variance is now estimated as -0.196, compared to an unadjusted difference of -0.266. In terms of the Public and Catholic sector population-averaged within-school variances, these are now estimated as 40.287 and 33.129 compared to 44.331 and 33.390 in Model 1 and so, when we adjust for SES, both residual variances reduce and the dispersion difference narrows. Each sector

variance is now interpreted as the expected variability in achievement among a group of same sector students with the same level of SES.

Refitting the model ignoring the level-1 variance function random effects v_{0j} (not presented) increases the DIC statistic by $46387 - 46373 = 14$ points and so the residual variance continues to vary randomly across schools, even after holding SES constant and accounting for the overall sector difference in dispersion.

Model 3

Another concern with the models specified so far is that the intake composition of a school may directly drive its subsequent achievement dispersion. We have already shown that schools which are more homogenous at intake (in terms of SES) have more homogenous achievement, but perhaps there is an additional contextual effect of social homogeneity whereby schools with more homogenous intakes prove easier to teach and so tend to produce lower achievement dispersion at any given level of SES. If this is the case then we might expect ignoring this effect will lead us to overestimate the true sector dispersion difference since Catholic schools tend to have more socially homogenous intakes than Public schools. Model 3 extends Model 2 by including the SD of SES as a predictor in the level-1 variance function. An additional concern is that the higher mean performance of Catholic students may lead them to be disproportionally affected by any ceiling effect of the achievement scale. We therefore also include mean SES in the model, expecting to see a negative coefficient should the ceiling effect be important.[3] We delay entering student-level SES into the level-1 variance function until Model 4.

$$\alpha_{0j} = \delta_{00} + \delta_{01}\text{SECTOR}_j + \delta_{02}(\text{MEAN SES})_j + \delta_{03}(\text{SD SES})_j + v_{0j},$$

The results (Table 3) show a worsening of the DIC statistic of $46375 - 46373 = 2$ points and while the coefficients of school mean SES and school SD of SES have the expected signs, neither have individually significant effects on dispersion.

Models 2 and 3 illustrate the process of adjusting for school differences in student background in terms of a single characteristic, SES. However, to obtain unbiased Type B mean effects we should include in the mean-function all student background characteristics and any school-level compositional variables predictive of math achievement whose means differ by school sector. Similarly, for the purpose of estimating unbiased Type B dispersion effects, we should additionally include in the mean-function all student background characteristics predictive of math achievement whose variances differ appreciably by school sector (even if their means do not) (Kim and Seltzer, 2011). Further, where school-level compositional variables are directly predictive of dispersion, these should be directly included in the level-1 variance function. Once an adequate model for Type B mean and dispersion effects has been specified, attention can then shift to exploring the role of different school policies and practices in explaining why schools in the Catholic sector appear, on average, more effective than those in the Public sector. For example, Raudenbush and Bryk (2002) note that students in Public schools pursue more differentiated course taking patterns than students in Catholic schools and this may contribute to the greater dispersion seen in the Public sector.

Model 4

While our illustration focuses on adjusting the sector mean and dispersion differences for selection into schools, it is also interesting to explore whether achievement dispersion varies within schools as a function of SES. For example, is the math achievement of low SES students more or less predictable than that of high SES students? Does any such relationship vary across schools and across the two sectors? Addressing these questions leads us to include school-mean centered SES in the level-1 variance function with a random-slope and a cross-level interaction with sector. The level-1 variance function is

$$\log\left(\sigma_{ij}^2\right) = \alpha_{0j} + \alpha_{1j}(\text{SES} - \text{MEAN SES})_{ij},$$

$$\alpha_{0j} = \delta_{00} + \delta_{01}\text{SECTOR}_j + v_{0j},$$

$$\alpha_{1j} = \delta_{10} + \delta_{11}\text{SECTOR}_j + v_{1j}.$$

where v_{0j} and v_{1j} are assumed bivariate normal with zero mean and constant covariance matrix. We also allow the four random effects to covary across the mean and level-1 variance function with covariances τ_{u0v0}, τ_{u1v0}, τ_{u0v1}, and τ_{u1v1}. Thus, each school has its own variance-function with its own intercept α_{0j} and its own slope α_{1j}. The intercept α_{0j} measures the log of the variation in math achievement for students with SES equal to their school mean SES. The slope α_{1j} measures how this variation changes as a function of student SES. The level-2 models predict the intercepts and slopes of these relationships by school sector.

Moving from Model 2 to 4 shows a decrease in the DIC statistic of $46373 - 46361 = 12$ points and so Model 4 (Table 3) provides the better fit to the data. Both the sector difference in intercepts and the sector difference in slopes are significant. The relationship

is positive and strong in the Public sector ($\delta_{10} = 0.103$) and negative and strong in the Catholic sector ($\delta_{10} + \delta_{11} = 0.103 - 0.239 = -0.136$). Figure 2 illustrates the substantial variability in intercepts and slopes by plotting the predicted level-1 variance functions.[4]

Refitting the current model ignoring the level-1 variance function random effects, v_{0j} and v_{1j}, (not presented) increases the DIC statistic by 46369 – 46361 = 8 points and results in the pattern of spurious precision suggested by the simulation study: the posterior SDs for δ_{01}, δ_{10} and δ_{11} decrease (by 17%, 8% and 9%, respectively) for little appreciable change in the posterior means.

One could in theory extend Model 4 in the same way we extended Model 1, by exploring heterogeneity of variance at level-2. Specifically we could model the 10 level-2 variances and correlations of the 4×4 level-2 covariance matrix as functions of the two school-level predictors. However, this leads to 10 functions with a total of 30 parameters, which is very high especially given that there are only 160 schools in the data. Modeling the level-2 covariance matrix as a function of level-2 predictors will typically prove most fruitful when the matrix is of a lower order or in studies where there are higher numbers of level-2 units.

6. Discussion

We have reviewed modeling heterogeneous variance-covariance components in two-level models. We described how to model the outcome and the residual variance jointly as separate random-coefficient models where the random effects in each model are allowed to covary and the resulting variance-covariance parameters can themselves be modeled as a function of the predictors. Restricted versions of this model can be fitted in various packages. We fitted the full model in the Stat-JR software. Supplementary materials

describing how to fit these models using Stat-JR are found at http://www.bristol.ac.uk/cmm/software/statjr/.

We showed via simulation that ignoring unexplained level-2 variation in the level-1 variances leads the level-1 variance-function regression coefficients to be estimated with spurious precision, especially regression coefficients relating to level-2 predictors. Researchers ignoring this issue therefore run the risk of making type I errors of inference about the sources of level-1 variance heterogeneity. A useful extension would be to explore a wider range of dataset sizes to establish general guidelines as to the number of units required at each level to perform variance modeling. Similarly, it will be interesting to reanalyze existing studies that model the level-1 variance as a function of predictors, but do not include random effects, to examine the practical differences this exclusion may make.

We illustrated the modelling extensions with a reanalysis of the classic HSB data. Schools are shown to vary in their achievement dispersion as well as their mean achievement, even after adjusting for the differing socioeconomic compositions of their student intakes. School sector predicts these differences with Catholic schools exhibiting higher mean achievement and lower achievement dispersion compared to Public schools. A naïve interpretation of this result is that Catholic schools narrow educational inequalities and Public schools widen them, but in this observational setting there are alternative potential explanations. First, there may be omitted student characteristics which tend to show lower variance in Catholic schools compared to Public schools. Second, school sector may be partially confounded with omitted school-level variables which are themselves predictive of achievement dispersion. With additional covariate information we could start to investigate these explanations. A third explanation is that Catholic students may be

disproportionally affected by a ceiling effect of the achievement scale, this could be explored by considering models for censored outcomes.

The principle of modeling variance-covariance parameters in two-level models as functions of the predictors and further random effects generalizes to more complex multilevel models in educational and behavioral research. For example, Leckie and Baird (2011) fit a two-way students-by-raters cross-classified model to students' essay scores. In this setting, the level-1 variance can be interpreted as measuring inconsistent scoring. Allowing this variance to vary randomly across raters would provide rater-specific estimates of rater inaccuracy allowing errant raters to be identified so that they can be retrained or removed from rating. As a second example, Leckie et al. (2012) fit a two-level schools-within-districts binomial logistic regression model for the proportion of low SES students in each school. In this setting, the level-1 variance can be interpreted as measuring social segregation within districts. Allowing this variance to vary randomly across districts would provide district-specific estimates of social segregation which could then be modeled by district-level predictors. A parallel extension could in principle be applied in the more general case of modelling multigroup segregation via multinomial logistic regression models (Leckie and Goldstein, 2015); the school-level covariance matrix would be allowed to vary randomly across districts. While these examples focus on cross-sectional settings in educational research, the modeling extensions are also relevant to longitudinal settings and other disciplines across the social and medical sciences.

It is prudent to end with some guidance and caution about these models. First, while we present a very general model for the level-1 variance function, it will often be the various sub-models that we present in our illustration which will prove most useful for many

researchers. Second, caution should be exercised when combining complex level-1 and level-2 random parts of the models as the resulting model may not always be identified. Cross-correlations and bivariate plots of the parameter MCMC chains should be checked to ensure there are no ridges in the posterior distribution (Cowles and Carlin, 1996). Fitting the model of interest separately to each cluster can also shed light on potential identification issues. More generally, different random-effect models may imply the same marginal model and so the modelling choices one makes must be carefully guided by substantive theory, not the fit of the data alone. Third, while inference in standard multilevel models has been shown to be relatively robust to violation of the multivariate normality assumption of the mean-function random-effects (McCulloch and Neuhaus, 2011), further simulation studies are needed to determine whether this finding also applies to the level-1 variance-function random-effects. Lastly, these models have proved difficult to fit by maximum likelihood estimation, while MCMC extends well to complex models with many random-effects. However, MCMC estimation is not without its own difficulties. In particular, one must specify prior distributions and in some settings small changes to the priors (especially for the level-2 covariance matrix) may have non-trivial effects on the posterior distributions (Browne and Draper, 2000; Seltzer et al., 1996). It is important that researchers test the sensitivity of their results to different choices of prior.

Notes

[1] Exponentiating δ_{01} gives a rate ratio interpretation, namely the ratio of the Catholic variance to the Public variance. Thus, the Catholic variance is $\exp(-0.266) = 0.766$ times the Public variance.

[2] This model provides the only convergence difficulties. Specifying diffuse priors for the kappa parameters sometimes led the different chains to converge on different posterior distributions. Specifying more informative Gaussian priors helps the chains converge to common stationary unimodal posterior distributions. We specify Normal($\hat{\boldsymbol{\kappa}}, \mathbf{I}$) where $\hat{\boldsymbol{\kappa}}$ is some plausible prior estimate for this vector (the intercept prior estimates are set to the Model 1 estimates suitably transformed, and the sector difference prior estimates are set to zero) and $\mathbf{I}$ is the identity matrix.

[3] We note that the sample mean and SD of SES are subject to a form of a measurement error because the ideal variables would be the mean and SD of SES of the population represented by the students.

[4] Note that distributional plots of unit-specific posterior means such as Figure 2 will typically be underdispersed (see Shen and Louis, 1998). One way to address this is to simply plot a new random sample of level-1 variance functions whose random effects are drawn from the estimated population distribution, but we have not pursued this here.

Appendix

Consider the following random-intercept model for the log of the level-1 variance

$$\log\left(\sigma_j^2\right) = \delta_{00} + v_{0j},$$

$$v_{0j} \sim \mathrm{N}(0, \tau_{v00}).$$

The level-1 variances are log-normally distributed

$$\sigma_j^2 \sim \log N\{E(\sigma_j^2), var(\sigma_j^2)\},$$

with mean (population-averaged level-1 variance)

$$E(\sigma_j^2) = \exp(\delta_{00} + \tfrac{1}{2}\tau_{v00}),$$

and variance (population-variance of the level-1 variances)

$$var(\sigma_j^2) = \{\exp(\tau_{v00}) - 1\}\exp(2\delta_{00} + \tau_{v00}).$$

Acknowledgements

This research was funded under the LEMMA3 project, a node of the UK Economic and Social Research Council's National Centre for Research Methods (grant number RES-576-25-0035). We are very grateful to the helpful comments made by the four reviewers.

References

Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics*, *36*, 332-339.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*, 117-128.

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)*, *149*, 1-43.

Browne, W. J. (2006). MCMC algorithms for constrained variance matrices. *Computational statistics & data analysis*, *50*, 1655-1677.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational statistics*, 15, 391-420.

Browne, W. J., Draper, D., Goldstein, H., & Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. *Computational statistics & data analysis*, *39*, 203-225.

Charlton, C.M.J., Michaelides, D.T., Parker, R.M.A., Cameron, B., Szmaragd, C., Yang, H., Zhang, Z., Frazer, A.J., Goldstein, H., Jones, K., Leckie, G., Moreau, L. and Browne, W.J. (2013). Stat-JR version 1.0. Centre for Multilevel Modelling, University of Bristol & Electronics and Computer Science, University of Southampton. URL http://www.bristol.ac.uk/cmm/software/statjr/.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association, *91*, 883-904.

Davidian, M., & Carroll, R. J. (1987). Variance function estimation. Journal of the American *Statistical Association*, *82*, 1079-1091.

Gilmour A, R., Gogel, B. J., Cullis, B. R., & Thompson, R. (2009). ASReml user guide release 3.0. VSN International, Hemel Hempstead.

Goldstein, H. (2011). Multilevel Statistical Models, 4th edition. Chichester, UK: Wiley

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A Multilevel Analysis of School Examination Results. *Oxford Review of Education*, *19*, 425-433.

Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159, 149-163.

Harvey, A. (1976). Estimating regression models with multiplication heteroskedasticity. *Econometrica*, *44*, 461-465.

Hedeker, D., & Nordgren R. (2013). MIXREGLS: A Fortran Program for Mixed-effects Location Scale Analysis. *Journal of Statistical Software*, *52*, 1-38.

Hedeker, D., & Mermelstein, R. J. (2012). Mood changes associated with smoking in adolescents: An application of a mixed-effects location scale model for longitudinal Ecological Momentary Assessment (EMA) data. In G. R. Hancock & J. Harring (Eds.), *Advances in Longitudinal Methods in the Social and Behavioral Sciences*.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An Application of a Mixed-Effects Location Scale Model for Analysis of Ecological Momentary Assessment (EMA) Data. *Biometrics*, *64*, 627-634.

Jahng, S. (2008). Multilevel models for intensive longitudinal data with heterogeneous error structure: covariance transformation and variance function models (Doctoral dissertation, University of Missouri-Columbia).

Kasim, R. M., & Raudenbush, S. W. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics*, *23*, 93-116.

Kim, J., & Choi, K. (2008). Closing the Gap: Modeling within-school variance heterogeneity in school effect studies. *Asia Pacific Education Review*, *9*, 206-220.

Kim, J., & Seltzer, M. (2011). Examining heterogeneity in residual variance to detect differential response to treatments. *Psychological methods*, *16*, 192-208.

Leckie, G. (2014). runmixregls - A Program to Run the MIXREGLS Mixed-effects Location Scale Software from within Stata. *Journal of Statistical Software, Code Snippet*, 1-41. *Forthcoming*.

Leckie, G., & Baird, J. A. (2011). Rater Effects on Essay Scoring: A Multilevel Analysis of Severity Drift, Central Tendency, and Rater Experience. *Journal of Educational Measurement*, *48*, 399-418.

Leckie, G., & Goldstein, H. (2015). A Multilevel Modelling Approach to Measuring Changing Patterns of Ethnic Composition and Segregation among London Secondary Schools, 2001-2010. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *Forthcoming*.

Leckie, G., Pillinger, R., Jones, K., & Goldstein, H. (2012). Multilevel Modeling of Social Segregation. *Journal of Educational and Behavioral Statistics*, *37*, 3-30.

Lee, Y., & Nelder, J. A (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B (Methodological), 58*, 619-678.

Lee, Y., & Nelder, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, *88*, 987-1006.

Lee, Y., & Nelder, J. A. (2006). Double hierarchical generalized linear models (with discussion). *Applied Statistics*, *55*, 139–185.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2006). Generalised linear models with random effects: unified analysis via h-likelihood. Chapman & Hall: London.

Littell, R. C. (2006). SAS for mixed models. SAS institute.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). The BUGS Book: A Practical Introduction to Bayesian Analysis. Chapman & Hall/CRC Texts in Statistical Science.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian Modelling Framework: Concepts, Structure, and Extensibility. Statistics and Computing, *10*, 325-337.

McCulloch, C. E., & Neuhaus, J. M. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, *67*, 270-279.

Noh, M., & Lee, Y. (2013). Package 'dhglm': Double Hierarchical Generalized Linear Models. http://CRAN.R-project.org/package=dhglm.

Payne, R. W., Murray, D. A., Harding, S. A., Baird, D. B., & Soutar, D. M. (2009). GenStat for Windows (12th Edition) Introduction. VSN International, Hemel Hempstead, UK.

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org/.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2009). MLwiN Version 2.1. Centre for Multilevel Modelling, University of Bristol, UK. URL http://www.mlwin.com.

Rast, P., Hofer, S. M., & Sparks, C. (2012). Modeling individual differences in within-person variation of negative and positive affect in a mixed effects location scale model using BUGS/JAGS. *Multivariate Behavioral Research*, *47*, 177-200.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*, 1–17.

Raudenbush, S. W., & Bryk, A. S. (1987). Examining correlates of diversity. *Journal of Educational and Behavioral Statistics*, *12*, 241-269.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and Data Analysis methods. 2nd edition. Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2012). HLM 7: Hierarchical linear and nonlinear modeling. Lincolnwood, IL: Scientific Software International, Inc. URL http://www.hlm.com.

Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*, 307-335.

SAS Institute Inc. (2013). Base SAS 9.3 Procedures Guide. Statistical Procedures, Second Edition. Cary, NC: SAS Institute Inc. URL http://www.sas.com/.

Seltzer, M. H., Wong, W. H., & Bryk, A. S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics*, *21*, 131-167.

Shen, W., & Louis, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*, 455-471.

Snijders, T. A. B., & Bosker, R. J. (2012). Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling. 2nd edition. Sage.

StataCorp. (2013). Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, *64*, 583-639.

Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 26, 209-232.

TABLE 1
Results of the parameter recovery simulation study.

Parameter		True value	Model A				Model B			
			Average Posterior Mean	SD of Posterior Mean	Average Posterior SD	Coverage	Average Posterior Mean	SD of Posterior Mean	Average Posterior SD	Coverage
Mean-function										
γ_{00}	Intercept	0	-0.003	0.051	0.050	95	-0.002	0.051	0.050	94
γ_{01}	Level-2 predictor (W_j)	0.2	0.201	0.052	0.051	93	0.200	0.052	0.051	94
γ_{10}	Level-1 predictor (X_{ij})	0.5	0.501	0.025	0.025	95	0.501	0.023	0.023	95
γ_{11}	Cross-level predictor ($W_j X_{ij}$)	-0.1	-0.099	0.025	0.024	95	-0.099	0.023	0.023	95
τ_{u00}	Intercept variance	0.1	0.103	0.027	0.027	94	0.103	0.027	0.027	95
Level-1 variance-function										
δ_{00}	Intercept	-0.611	-0.519	0.078	0.042	42	-0.611	0.075	0.076	94
δ_{01}	Level-2 predictor (W_j)	0.1	0.105	0.082	0.043	71	0.104	0.080	0.077	94
δ_{10}	Level-1 predictor (X_{ij})	0.1	0.099	0.056	0.042	86	0.101	0.048	0.047	94
δ_{11}	Cross-level predictor ($W_j X_{ij}$)	-0.1	-0.102	0.057	0.044	86	-0.101	0.049	0.048	95
τ_{v00}	Intercept variance	0.2	–	–	–	–	0.207	0.060	0.063	95

TABLE 2
Results for Model 1 and 2

Parameter		Model 1 Posterior Mean	Model 1 Posterior SD	Model 2 Posterior Mean	Model 2 Posterior SD
Mean-function					
γ_{00}	Intercept	11.379	0.299	12.101	0.204
γ_{01}	Sector	2.829	0.442	1.258	0.310
γ_{02}	Mean SES	–	–	5.281	0.370
γ_{10}	SES deviation	–	–	2.944	0.170
γ_{11}	Sector × SES deviation	–	–	-1.664	0.250
γ_{12}	Mean SES × SES deviation	–	–	1.045	0.313
τ_{u00}	Intercept variance	6.817	0.889	2.521	0.384
τ_{u11}	Slope variance	–	–	0.330	0.143
ρ_{u01}	Intercept-Slope correlation	–	–	0.054	0.213
Level-1 variance-function					
δ_{00}	Intercept	3.783	0.028	3.687	0.028
δ_{01}	Sector	-0.266	0.040	-0.196	0.040
τ_{v00}	Intercept variance	0.016	0.006	0.017	0.006
Cross-function correlation					
ρ_{u0v0}	Intercept-Intercept correlation	-0.279	0.173	-0.341	0.166
ρ_{u1v0}	Slope-Intercept correlation	–	–	0.314	0.243
DIC		46817		46373	

TABLE 3
Results for Model 3 and 4

Parameter		Model 3 Posterior Mean	Model 3 Posterior SD	Model 4 Posterior Mean	Model 4 Posterior SD
Mean-function					
γ_{00}	Intercept	12.103	0.205	12.088	0.205
γ_{01}	Sector	1.241	0.313	1.282	0.315
γ_{02}	Mean SES	5.338	0.383	5.162	0.382
γ_{10}	SES deviation	2.937	0.170	2.959	0.170
γ_{11}	Sector × SES deviation	-1.656	0.252	-1.678	0.252
γ_{12}	Mean SES × SES deviation	1.005	0.316	0.958	0.316
τ_{u00}	Intercept variance	2.515	0.384	2.550	0.382
τ_{u11}	Slope variance	0.325	0.144	0.340	0.136
ρ_{u01}	Intercept-Slope correlation	0.054	0.212	0.095	0.204
Level-1 variance-function					
δ_{00}	Intercept	3.676	0.030	3.683	0.028
δ_{01}	Sector	-0.171	0.046	-0.197	0.040
δ_{02}	Mean SES	-0.062	0.054	–	–
δ_{03}	SD SES	0.100	0.221	–	–
δ_{10}	SES deviation	–	–	0.103	0.039
δ_{11}	Sector × SES deviation	–	–	-0.239	0.058
τ_{v00}	Intercept variance	0.016	0.006	0.016	0.005
τ_{v11}	Slope variance	–	–	0.016	0.007
ρ_{v01}	Intercept-Slope correlation	–	–	0.071	0.259
Cross-function correlation					
ρ_{u0v0}	Intercept-Intercept correlation	-0.346	0.164	-0.319	0.163
ρ_{u1v0}	Slope-Intercept correlation	0.307	0.245	0.217	0.242
ρ_{u0v1}	Intercept-Slope correlation	–	–	-0.352	0.198
ρ_{u1v1}	Slope-Slope correlation	–	–	-0.225	0.254
DIC		46375		46361	

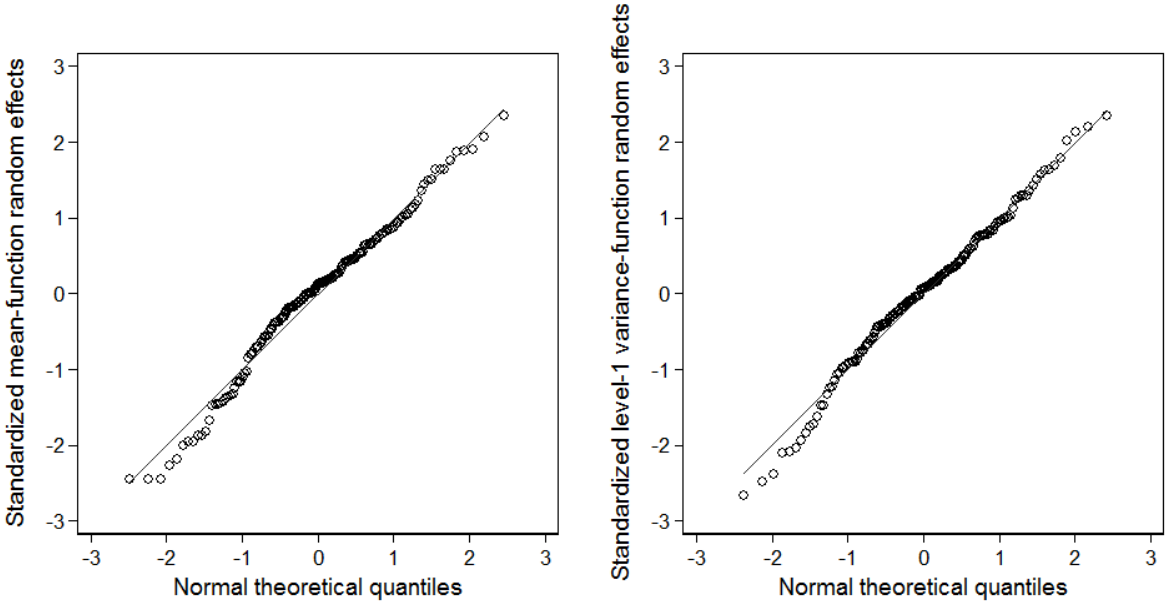
Standardized mean-function random effects
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Normal theoretical quantiles
Standardized level-1 variance-function random effects
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Normal theoretical quantiles

FIGURE 1.

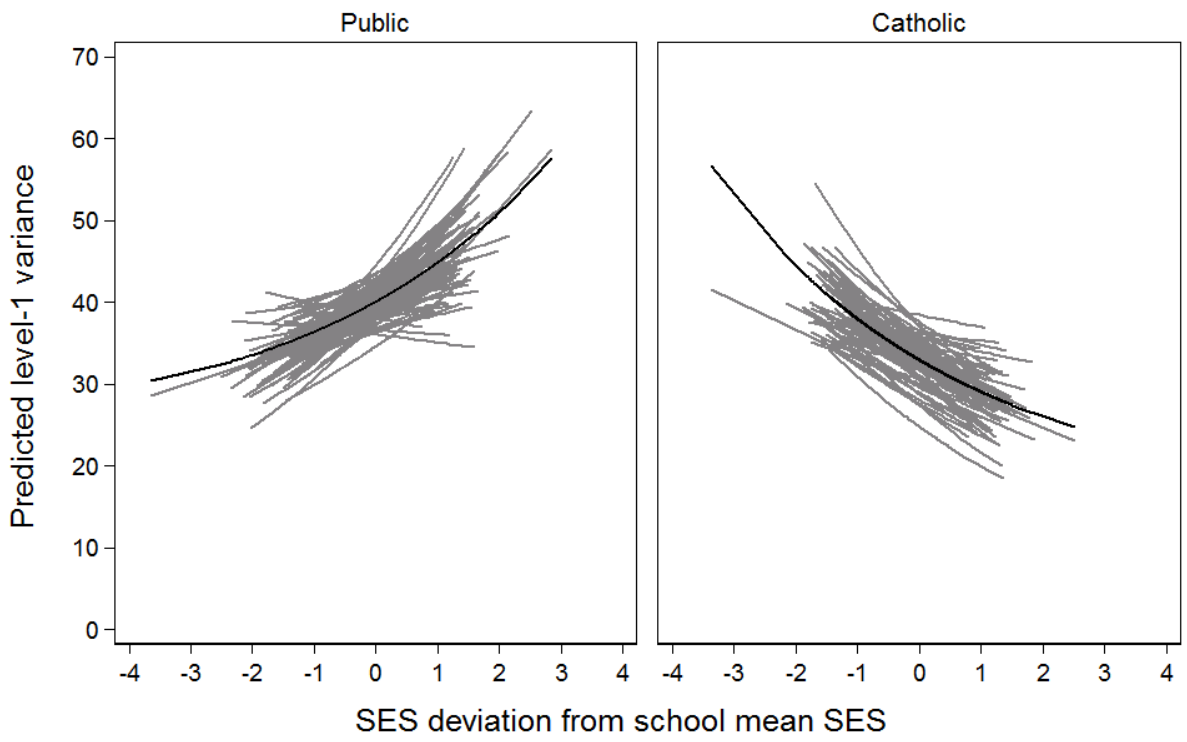
Public
Catholic
Predicted level-1 variance
70
60
50
40
30
20
10
0
-4
-3
-2
-1
0
1
2
3
4
SES deviation from school mean SES

FIGURE 2.

Captions

FIGURE 1. *Model 1 quantile-quantile plots of the standardized mean-function random-intercept effects (left panel) and variance-function random-intercept effects (right panel).*

FIGURE 2. *Model 4 predicted school achievement dispersion as a function of student SES, plotted separately within Public and Catholic sectors.*